Behavior Research Methods (2024) 56:3009–3022
https://doi.org/10.3758/s13428-023-02290-3

ORIGINAL MANUSCRIPT

Chasing consistency: On the measurement error in self-reported affect in experiments

Niels Vanhasbroeck[1] · Sophie Vanbelle[2] · Agnes Moors[1,3] · Wolf Vanpaemel[1] · Francis Tuerlinckx[1]

Accepted: 6 November 2023 / Published online: 22 November 2023

Abstract

How feelings change over time is a central topic in emotion research. To study these affective fluctuations, researchers often ask participants to repeatedly indicate how they feel on a self-report rating scale. Despite widespread recognition that this kind of data is subject to measurement error, the extent of this error remains an open question. Complementing many daily-life studies, this study aimed to investigate this question in an experimental setting. In such a setting, multiple trials follow each other at a fast pace, forcing experimenters to use a limited number of questions to measure affect during each trial. A total of 1398 participants completed a probabilistic reward task in which they were unknowingly presented with the same string of outcomes multiple times throughout the study. This allowed us to assess the test–retest consistency of their affective responses to the rating scales under investigation. We then compared these consistencies across different types of rating scales in hopes of finding out whether a given type of scale led to a greater consistency of affective measurements. Overall, we found moderate to good consistency of the affective measurements. Surprisingly, however, we found no differences in consistency across rating scales, which suggests that the specific rating scale that is used does not influence the measurement consistency.

Keywords Affect dynamics · Consistency · Experiment · Measurement · Reliability

How we feel plays an important part in our lives. It is therefore not surprising that much research has gone into how our feelings, or *affective states*, change over time. Within this field—often dubbed *affect dynamics*—researchers have investigated the way in which affect changes, the reasons why affect changes, and whether there are any systematic individual differences in both aspects. Regarding the way in which affect changes, investigators have defined several guiding principles for understanding affective fluctuations over time, such as regulation towards a baseline affective state (e.g., Kuppens & Verduyn, 2017; for model-based accounts, see Driver & Voelkle, 2018; Loossens et al., 2020; Vanhasbroeck et al., 2021). With regard to the reasons why affect changes, many researchers have connected changes in affect to contextual factors, such as experimental stimuli (Eldar & Niv, 2015; Rutledge et al., 2014; Vanhasbroeck et al., 2022) or daily-life events (Burns & Ma, 2015; Dejonckheere et al., 2021; Villano et al., 2020), and several others have proposed theories to explain the connection (Cunningham et al., 2013; Frijda, 2007; Moors et al., 2021). Finally, research has identified individual differences in affect dynamics, with regard to both personality and psychopathology (Bonsall et al., 2012; Huys et al., 2013; Smillie et al., 2013; Trull et al., 2015).

Typically, investigators use self-report measures to study affect dynamics. This requires participants to indicate their feelings on a rating scale paired with a question such as "How happy are you feeling right now?" Self-report rating scales are an intuitive choice for measuring affect. However, the reliance of the field on self-report measures of affect requires that these measures possess good psychometric qualities.

One such quality is that measures of affect are not very susceptible to noise, also referred to as *measurement error*. Within the field of affect dynamics, measurement error has received increased attention after researchers were confronted with several counterintuitive results. For example,

✉ Niels Vanhasbroeck
niels.vanhasbroeck@kuleuven.be

[1] Research Group of Quantitative Psychology and Individual Differences, KU Leuven, Leuven, Belgium

[2] Department of Methodology and Statistics, Maastricht University, Maastricht, the Netherlands

[3] Center for Social and Cultural Psychology, KU Leuven, Leuven, Belgium

Dejonckheere et al. (2019) found no evidence for the added value of dynamical features of affect for the prediction of psychopathological symptoms over and above the mean and variance, questioning the importance of studying these dynamical features in relation to psychopathology. This result has since been corroborated by counterintuitive results in the relation of affect to personality (e.g., Kalokerinos et al., 2020; Wendt et al., 2020) and in the description of affect using computational modeling (Bulteel et al., 2018; Loossens et al., 2021). It is argued that one of the culprits of these findings is measurement error (Dejonckheere & Mestdagh, 2021), stressing the importance of studying, identifying, and controlling for error in our measurements.

Unfortunately, an important obstacle in studying measurement error in affect is that the primary variable of interest changes continuously over time. Despite this difficulty, several authors have attempted to quantify measurement error, either through taking advantage of the study's design (Dejonckheere et al., 2022; Eisele et al., 2021), through the decomposition of variance (Haney et al., 2023; Scott et al., 2020; Wilhelm & Schoebi, 2007), or through the use of advanced modeling techniques (Schuurman et al., 2015). Results of these studies have been diverse: Across studies, 10% to 50% of the observed variation in affect has been attributed to error. This means that, while we pick up on some systematic variation in affect, measurement error is not negligible.

Importantly, these previous studies all focused on assessing measurement error in daily life. While daily-life studies enjoy great ecological validity, these studies suffer from having only limited information on the stimuli or events that influence a participant's affective states throughout the study. This limits the extent to which researchers are able to assess measurement error independently of changes in affect, making these studies more susceptible to the under- or overestimation of measurement error (Scott et al., 2020).

As an alternative, experimental laboratory studies allow for more fine-grained control of these stimuli or events. This increased control allows us to disentangle affect from measurement error, ultimately providing us with a more accurate estimate of measurement error. With this in mind, we set up an experimental task to assess and compare measurement error for several types of rating scales fit for lab-based research. We additionally investigated how users experienced these different scales.

To assess measurement error, we computed each scale's reliability, defined as the extent to which a variable is measured consistently across replications (Lord et al., 1968). To estimate reliability, we relied on generalizability theory. This theory is an extension of classical test theory in which multiple sources of variation can be identified as being either systematic or unsystematic (Brennan, 2001). To identify these different sources of variation, one defines a variance decomposition model in which all suspected sources of variation are accounted for in a similar way as the typical analysis of variance (ANOVA). Importantly, both the identified sources of variation and their classification as either systematic or unsystematic depend heavily on the researcher's design and interest. Here lies a major strength of generalizability theory: It offers a rich and flexible framework for investigating one's measures.

Importantly, the variance components can be used to compute a measure of reliability of one's measures. This is achieved by dividing the systematic variation in interest by a sum of both systematic and unsystematic variance components. In general, researchers distinguish between two types of reliability coefficients. The first coefficient is often referred to as *relative reliability* or *consistency* and assesses the extent to which measured responses show a similar pattern across measurement occasions. The second coefficient is regularly referred to as *absolute reliability* or *agreement* and represents the extent to which measured responses match each other exactly across iterations.

Given their definitions, it should not come as a surprise that both relative and absolute reliability have an intricate relationship with measurement error, so that high reliability values go together with little measurement error and vice versa. Indeed, because of this relationship, these measures are regularly used to determine the extent of the measurement error in one's ratings, just as we do in this study (see, e.g., Haney et al., 2023; Lucas & Donnellan, 2012; see also Lord et al., 1968; McGraw & Wong, 1996).

Note that we will use the term "consistency" to refer to reliability throughout this article. The reasons for this choice are twofold. First, it explicitly cuts ties with the often-implied link between reliability and classical test theory. Second, it allows us to refer to both relative and absolute reliability with a single term.

It is important to note that our study should be taken as complementary to daily-life studies, and not as a validation of their results. At present, it is unclear whether results obtained in daily life generalize to laboratory studies and vice versa, as both kinds of designs pose different limitations to the measurements one can obtain. For example, while multiple items can be used to measure the same construct within daily-life studies, this is more difficult to achieve within an intensive laboratory setting, requiring experimenters to rely on a minimum number of items to assess affect (e.g., single-item rating scales, Rutledge et al., 2014; Vanhasbroeck et al., 2022). This is especially true in multi-trial experiments in which researchers are required to measure affect at a high frequency. Researchers might also use stimuli that are too simple to elicit multidimensional responses—for example, as regards monetary decision outcomes. The results of this study thus might not generalize to daily-life studies of affect dynamics.

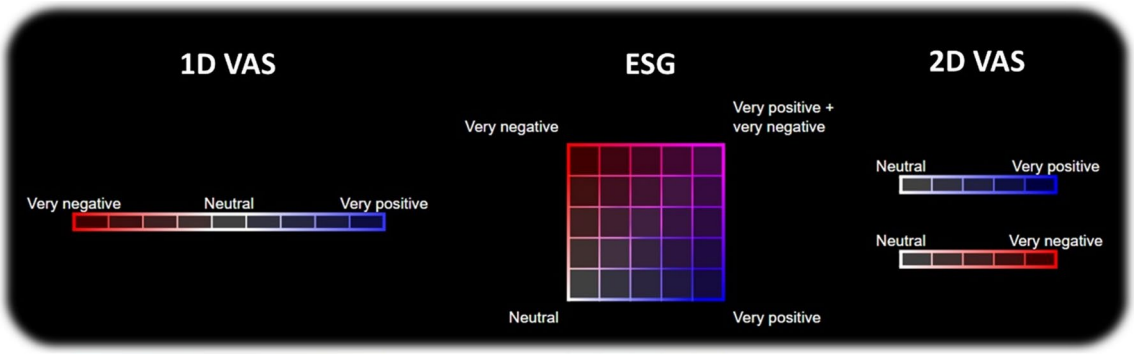

Fig. 1 Visualization of the three rating scale formats

Method

This study was preregistered on the Open Science Framework on January 19, 2023 (https://osf.io/sytrn). It also comes with open materials that are accessible on Gitlab[1]. These materials include code to run the experiment, code to run the analyses, the raw and preprocessed data, and the results. In this paper, we report how we determined our sample size, all exclusions, all manipulations, and all measures that we used in the study.

Interested readers can find the *Supplementary Materials* via the following link: https://osf.io/ce87p.

Participants

We recruited the 1416 participants through the online data collection platform Prolific (https://www.prolific.co/). As detailed in our preregistration, this sample size was based on a simulation study in which we determined the sample size, number of sequences, and number of trials per sequence to use in our experiment while accounting for the total cost of this study. From these initial 1416 participants, 18 were excluded from the analyses. Ten of these individuals were excluded because they answered "no" to the question of whether they believed we should use their data, which was asked at the end of the study. An additional eight individuals were excluded because their data were either not (four participants) or not fully (four participants) registered in the database, all for currently unknown reasons[2]. The final sample size thus consisted of 1398 participants. Most of these were men (60% male; 38% female; 2% other) and people who enjoyed a higher education (26% high school, 45% bachelor, 23% master, 6% other). Participants were on average 34 years old, showing a wide range of ages ($SD = 13$, range = [18,76]).

As participants would complete a probabilistic reward task, they were told that they would receive their total earnings as the reward for participation. Unbeknownst to them, this total was predetermined, so that everyone received £5 after successfully completing the experiment, irrespective of their behavior during the task. On average, participants spent 23min 59s on task (SD = 9min 6s; range = [11min 21s, 79min 39s]).

Materials

Rating scales

In this study, each participant was assigned one of six rating scales on which they had to indicate their momentary affect during the experiment. Each of these rating scales consisted of a given format (i.e., visual layout) and a given continuity (i.e., whether participants could provide continuous or discrete responses).

Regarding the scale format, we used three different types of scales in this study, all of which are visualized in Fig. 1. First, we used a visual analogue scale (VAS) that measured valence from very negative (0) to very positive (1). This kind of scale is often used in experimental studies, as it allows the quick measurement of a single construct of interest. However, a single VAS item is not adequate when one is interested in more than one psychological construct, for example, positive and negative affect (PA and NA, respectively) when these are considered to be independent constructs. For this, researchers have to turn to other rating scales, such as the evaluative space grid (ESG), which was used as the second format in our study. The ESG allows for the simultaneous assessment of PA and NA, ranging from neutral (0) to either very positive or very negative (1). This scale was developed to allow the measurement of PA and NA without requiring participants to answer questions on two different VAS scales (Larsen et al., 2008). However, the consistency of the ESG has not yet been compared to

[1] https://gitlab.kuleuven.be/ppw-okpiv/researchers/u0123135/affective-consistency

[2] The absence of these data was double-checked. On the Prolific page of our study, we found that a total of 1416 individuals completed the experiment. In the database, however, we received data for only 1412 of these, four of which were only partial.

the consistency of the measurements from two separate VAS scales. Therefore, we also included a VAS equivalent of the ESG as a third format, with two separate VASs measuring PA and NA in a similar way as the ESG (for a similar approach, see Asutay et al., 2021). To differentiate between the two kinds of VAS scales, we will use the acronyms *1D VAS* to refer to the one measuring bipolar valence and *2D VAS* to refer to the one measuring PA and NA.

With regard to continuity, each of the formats above could allow either discrete or continuous responses. When the rating scale was discrete, it functioned as a typical Likert scale in which participants were only able to submit a response that fell in the center of the cells of the scale (see again Fig. 1). When the rating scale was continuous, participants were allowed to respond anywhere on the rating scale. In the latter case, the cells of the rating scales were still visible to help participants in interpreting the scales.

When the rating scales were discrete, we allowed nine discrete response options for the 1D VAS. For the ESG and 2D VAS, participants could click in five cells along the PA and NA dimensions (including a "neutral" option), leading to nine discrete response options along the axes of the scale (going from "very negative" to "neutral" to "very positive"). Controlling this number made comparison between the three formats more straightforward.

Crossing the levels of each factor resulted in a total of 3 (format) × 2 (continuity) = 6 rating scales. Each participant was assigned one of these rating scales based on the time at which they started the study.[3] Due to the nature of the assignment, the sample size across conditions was somewhat unbalanced, as visualized in Table 1.

Table 1 Number of individuals that used a given rating scale

	Format		
Continuity	1D VAS	ESG	2D VAS
Discrete	233	219	238
Continuous	226	250	232

Manipulation checks, motivation, ease, and accuracy

To investigate how participants experienced the experiment and the rating scale they used, we asked them to fill out a questionnaire at the end of the study (see Table 2). With this questionnaire, we wanted to measure the following things. First, we wanted to know whether the stimuli used in the paradigm affected how participants really felt. Given the importance of eliciting real feelings in order to be able to measure them with the rating scales, this dimension was referred to as a *Manipulation check*. Second, we measured the participants' motivation to continue the study, which we labeled *Motivation* in the table. Third, we measured the ease with which participants could interpret and use the rating scale they were given, which we referred to as *Ease*. The final variable we measured was *Accuracy*, or the extent to which participants were able to accurately report on their feelings throughout the experiment. This variable required participants to estimate how closely their feelings related to the responses they made on the rating scales. Of all questionnaire variables, *Ease* and *Accuracy* were of particular interest given that they measured other desired scale characteristics. Consequently, these variables are given greater attention throughout this article.

Participants had to answer each of the questions in Table 2 on a seven-point Likert scale from "strongly disagree" to "strongly agree." The order in which questions are shown in this table is also the order in which the questions were shown to the participants. All of these questions were available at the same time.

Depressive symptoms

Finally, we measured depressive symptoms with the Center for Epidemiological Studies Depression scale (CES-D, Radloff, 1977). As mentioned in our preregistration, this was an exploratory variable meant for future use. We do not discuss the CES-D in the remainder of this article.

Procedure

After giving their consent, participants received instructions on how to complete the experiment and how to use the assigned scale. Then, participants completed a self-paced probabilistic reward task developed in our lab (see Vanhasbroeck et al., 2022). In this task, participants were shown four doors, each of which hid either a monetary win or loss. On each trial, participants had to choose one of these doors to receive the concealed monetary outcome and add it to their total. When a door was chosen, this choice remained visible on screen for 500ms, after which the door opened and showed the monetary outcome for 2s.

At the end of each trial, participants had to report their affective state on the assigned rating scale. The temporary responses of participants were tracked by placing a red dot at the clicked location. If a participant wished to proceed to the next trial, they had to confirm their response by pressing the spacebar. To avoid imposing dependencies between trials, the red dot was removed at the start of each trial.

[3] This assignment was performed in the following way: We generated a number from the exact time at which the participant clicked our study link, where time was put in the format HH:MM:SS.MS. Then we computed the number's remainder after dividing by six. The result of this procedure was one of six possible numbers, which then determined the condition to which the participant was assigned.

Table 2 Questions used to assess participants' experience of the experiment and the affect measures

Intended variable	Question
Manipulation check	*The experiment elicited positive feelings.*
	The experiment elicited negative feelings.
	I felt indifferent about the experiment.
	Winning or losing money did not have any effect on me.
Motivation	*I was motivated to finish the experiment.*
	The experiment was boring.
	The experiment was frustrating.
	I enjoyed the experiment.
	The experiment was over sooner than I thought.
Ease of use	*The emotion measure was easy to use.*
	I quickly understood how to use the emotion measure.
	It was difficult to report on my feelings.
	The emotion measure was confusing to use.
Accuracy	*I was able to accurately describe my feelings with the emotion measure.*
	When my feelings changed, I could describe these changes accurately with the emotion measure.
	My responses on the emotion measure conveyed how I really felt during the experiment.
	My feelings could not be adequately captured by the emotion measure.

Importantly, participants were told that they had an equal probability of winning and losing on each trial. In reality, however, their monetary outcomes were predetermined and fixed in such a way that the same sequence of 28 outcomes was repeated five times throughout the experiment (for an illustration, see Fig. 2). This within-person manipulation allowed us to assess the test–retest consistency for the rating scale a participant used, as will be detailed in the *Analyses* section.

The monetary outcomes were generated in the following way. First, 14 wins and 14 losses were drawn from the set {0.10,0.15,0.20,0.25,0.30,0.35,0.40,0.45,0.50} (losses were obtained by multiplying the drawn number by −1). Then, the prevalence of each discrete outcome was manipulated in such a way that the sum of all monetary outcomes would be equal to £1. This manipulation did not change the relative occurrence of wins and losses, that is, participants went through an equal number of wins and losses. Lastly, these outcomes were randomized within-person and then repeated five times. We will refer to each of such repetitions as a sequence in the remainder of this paper.

At the end of the study, participants were asked to fill out the two questionnaires, starting with the questionnaire about their experiences and then moving on to the CES-D. Only after filling out the first questionnaire could participants move on to the second one. Questions were always shown in the same order.

When both questionnaires were filled out, participants were debriefed. They were furthermore asked to answer a

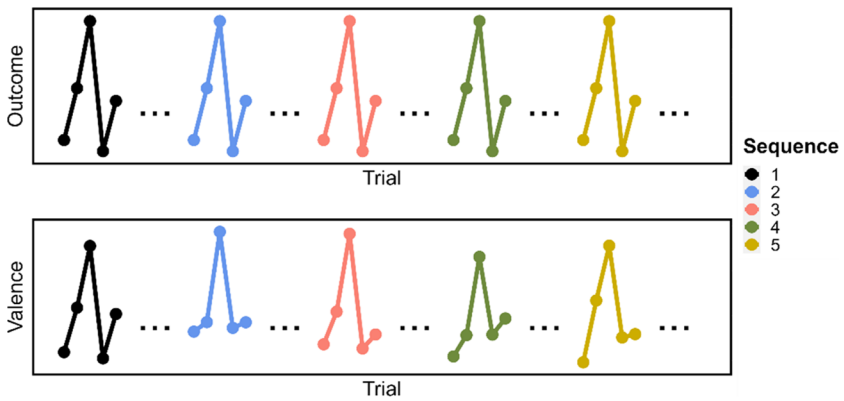

Fig. 2 Visualization of the main manipulation within this experiment. In the top plot, the monetary outcomes of the experiment are plotted across time, where each of the five sequences is given a specific color. In the bottom plot, the resulting affective time series is shown. As one may notice, the monetary outcomes were the same across sequences and were therefore expected to elicit similar affective states across sequences

diligence question, asking whether we should use their data ("In your honest opinion, should we use your data?"). Participants were given the option to respond with a simple "yes" or "no" and were provided with a text box in which they could optionally explain their choice.

Analyses

The analyses can be divided into two parts: analyses related to the consistency of the rating scales and analyses related to the questionnaire about the participants' experiences.

Consistency

Given that the same sequence of monetary outcomes was repeated multiple times throughout the experiment, we were able to estimate the consistency of a rating scale based on participants' affective responses on each of these outcomes. As a measure of consistency, we estimated several intraclass correlation coefficients (ICCs) using the data. The ICC is a measure of consistency that is often used to estimate inter-rater reliability, but it is equally well suited for estimating the test–retest consistency within the context of this experiment (Matheson, 2019; McGraw & Wong, 1996; Polit, 2014; ten Hove et al., 2022a). It is defined as the systematic variance divided by the total variance, thus communicating the proportion of systematic variation in one's variable of interest.

To compute an ICC, the investigator has to define a variance decomposition model in which they differentiate the systematic from the unsystematic variance. After having identified the sources of variation that are deemed to be interesting, the investigator pits the systematic variances against the variances that they assume to reflect error.

To estimate the ICCs, we used two different decomposition models, each of which will be explained in detail below. Each decomposition model—and their associated ICCs—was estimated within a Bayesian framework using the *rstan* package in R (R Core Team, 2021; Stan Development Team, 2022). This package uses Markov chain Monte Carlo to approximate the posterior distribution of the parameters specified in our decomposition models. This method requires specifying the number of approximations (i.e., the number of chains) and the number of samples that should be drawn from the posterior distribution (i.e., the number of iterations). With regard to the latter, one must also specify the number of iterations used to initialize the model (i.e., the burn-in period) to allow the chains to converge on the posterior distribution before effectively sampling from it. For our analyses, we specified the number of chains to be four, the number of burn-in samples to be 1000, and the number of total iterations to be 2000. To check the convergence of the samples, we followed suggestions from the Bayesian literature by confirming whether the $\hat{R}$ for all parameters was lower than 1.1 and by verifying that no divergent transitions occurred during the sampling process (Kruschke, 2015).

Regarding the priors of our Bayesian model, we used lognormal priors with a mean of 0 and variance of 1 for the estimated standard deviations in the decomposition models. Additionally, for the overall mean in the decomposition, we used a uniform prior with its bounds set as the minimum and maximum of the observations. No explicit prior was used for the ICCs because each of its components was already assigned one.

Preprocessing affect Before explaining how we computed the ICCs, let us first explain how we treated the affective variables. As one might have noticed, the rating scales were either one-dimensional (1D VAS) or two-dimensional (ESG and 2D VAS). Given that we estimated ICCs for each affective dimension separately, this complicated the comparison between rating scale formats. To solve this issue, we computed a derived valence score for both the ESG and the 2D VAS. More specifically, we define valence at time t for the two-dimensional scales as:

$$valence_t = \frac{1}{2}\left(PA_t - NA_t + 1\right)$$

which ranges from 0 to 1, just like the ratings for the 1D VAS. The consistency of these valence scores was compared to the consistency of valence as measured by the 1D VAS.

Consistency of affect The first decomposition model extracted the most important sources of variation in the affective ratings directly. For this decomposition to make sense, one should notice that the main sources of affective variation come from person-specific effects (individual differences) and from the monetary outcomes the participants were confronted with (affective reactivity). More specifically, each individual was asked to rate their affective state in response to several repetitions of the same monetary outcomes, so that the affective ratings could be decomposed in the following way:

$$y_{por} = \mu + \mu_p + \mu_o + \mu_{po} + \epsilon_{por} \tag{1}$$

where we specify the person and outcome indices as $p \in \{1, \ldots P\}$ and $o \in \{1, \ldots O\}$. We furthermore denote the residuals with the index $r \in \{1, \ldots R\}$. Importantly, we distinguished the interaction term μ_{po} from the error term ϵ_{por} in this equation.

By assuming that each of the components within Eq. 1 is normally distributed, we estimate the variances associated with each component as:

$$\begin{aligned} \mu_p &\sim N\left(0, \sigma^2_{person}\right) \\ \mu_o &\sim N\left(0, \sigma^2_{outcome}\right) \\ \mu_{po} &\sim N\left(0, \sigma^2_{person \times outcome}\right) \\ \epsilon_{por} &\sim N\left(0, \sigma^2_{residual}\right) \end{aligned} \tag{2}$$

Here, σ^2_{person} represents the systematic variation across individuals (i.e., overall interindividual differences), $\sigma^2_{outcome}$ represents the systematic variation across monetary outcomes (i.e., the affective variation induced by the outcomes), and $\sigma^2_{person \times outcome}$ represents the variation attributed to interindividual differences in one's responses to the monetary outcomes.

Based on this decomposition, we defined the following two ICCs of interest:

$$\begin{aligned} ICC_{person}(C,1) &= \frac{\sigma^2_{person}}{\sigma^2_{person}+\sigma^2_{person\times outcome}+\sigma^2_{residual}} \\ ICC_{outcome}(C,1) &= \frac{\sigma^2_{outcome}}{\sigma^2_{outcome}+\sigma^2_{person\times outcome}+\sigma^2_{residual}} \end{aligned} \tag{3}$$

The $ICC_{person}(C,1)$ reflects the between-person consistency, which measures how consistently we can tell one person apart from another. In other words, this ICC measures how consistently we can pick up on interindividual differences in the mean affective state. The $ICC_{outcome}(C,1)$ reflects the consistency with which we can tell the monetary outcomes apart in the affective ratings. It is thus a measure of how consistently a given monetary outcome elicits a given affective response across participants.

Both ICCs use the specification $(C,1)$, which means that they are aimed at measuring the consistency of a single rating. This means that this decomposition operates on the single item level, assessing how consistently participants indicate their affective state during a single trial. For the sake of readability, we will abbreviate these ICCs as ICC_{person} and $ICC_{outcome}$, dropping the specification $(C,1)$.

Agreement of model parameters While it is important to know how consistently we measure affect, it is equally important to investigate whether analyses based on those data will lead to the same results. We therefore examined the extent to which estimates of model parameters remained the same when the same experimental manipulations were being carried out. Theoretically speaking, these model parameters should be stable and they should not vary too much from occasion to occasion.

For this, we propose a second decomposition model in which we did not decompose affect, but we rather decomposed the parameter estimates that we obtained through analyzing affect. For this, we first estimated several individual- and sequence-specific autoregressive models with a moderated intercept (Adolf et al., 2017), defined as[4]:

$$\begin{aligned} y_{pst} &= f\left(x_{pst}\right) + \beta_{2,ps} y_{ps,t-1} + \upsilon_{pst} \\ f\left(x_{pst}\right) &= \beta_{0,ps} + \beta_{1,ps} x_{pst} \end{aligned} \tag{4}$$

[4] Note that in the following set of equations, we use commas in the subscripts to distinguish denotations (e.g., parameter numbers and names for the variance components) from running indices (e.g., person, sequence, and time).

where the indices $p \in \{1,\ldots P\}$, $s \in \{1,\ldots S\}$, and $t \in \{1,\ldots T\}$ specify the person, sequence, and trial within a sequence, respectively. The variable x_{pst} furthermore contains the monetary outcome a person p received at trial t within sequence s. Importantly, the intercept $f(x_{pst})$ of this autoregressive model depends on the monetary outcomes x_{pst} in a linear way, as captured by the parameter $\beta_{1,ps}$. In other words, the intercept increases with an increasing value of the monetary outcome.

To find out how well these estimates agree across sequences, we define the parameter estimates as $\beta_{j,ps}$ with $j \in \{0,1,2\}$ denoting the parameter of interest and decompose them as:

$$\beta_{j,ps} = \mu_j + \mu_{j,p} + \mu_{j,s} + \epsilon_{j,ps} \tag{5}$$

of which the components were distributed according to:

$$\begin{aligned} \mu_{j,p} &\sim N\left(0, \sigma^2_{j,person}\right) \\ \mu_{j,s} &\sim N\left(0, \sigma^2_{j,sequence}\right) \\ \epsilon_{j,ps} &\sim N\left(0, \sigma^2_{j,residual}\right) \end{aligned} \tag{6}$$

In these equations, $\sigma^2_{j,person}$ informs us on the systematic interindividual differences in the estimates for parameter j, and $\sigma^2_{j,sequence}$ represents the systematic effect of the sequence on the same estimate.

Based on this decomposition, we defined a second set of ICCs as:

$$\begin{aligned} ICC_{j,person}(A,1) &= \frac{\sigma^2_{j,person}}{\sigma^2_{j,total}} \\ &= \frac{\sigma^2_{j,person}}{\sigma^2_{j,person}+\sigma^2_{j,sequence}+\sigma^2_{j,residual}} \end{aligned} \tag{7}$$

Here, the $ICC_{j,\,person}(A,1)$ reflects the absolute agreement of the parameter estimates for each individual, or how closely parameter estimates resemble each other across sequences of the experiment. More specifically, $ICC_{0,\,person}(A,1)$, $ICC_{1,\,person}(A,1)$, and $ICC_{2,\,person}(A,1)$ measure the agreement of the intercept, the slope of the monetary outcomes, and the autoregressive effect across sequences. Importantly, the 1 in the definition of the ICC for this decomposition does not refer to the single item, but to a single sequence. To put it differently, this decomposition assesses how consistently participants indicate their affective states across a single sequence of trials, thus functioning on a different level of generalization than the first decomposition. We will abbreviate the ICCs for each parameter as $ICC_{j,\,person}$, dropping the specification $(A,1)$.

Interpretation Interpreting the size of the ICCs is always somewhat subjective, given that it depends on one's research question (Liljequist et al., 2019; McGraw & Wong, 1996). For example, one may expect a smaller amount of measurement

error in measurements of body temperature compared to measurements of self-reported pain levels, and thus hold thermometers to a different standard than pain questionnaires with regard to the interpretation of their consistency.

In this study, we base interpretation on the guidelines provided by Koo and Li (2016) for interpreting the ICCs. These guidelines classify ICCs as "poor," "moderate," "good," or "excellent" when this ICC is lower than 0.50, between 0.50 and 0.75, between 0.75 and 0.90, or higher than 0.90, respectively.

Comparing ICCs In this study, we not only estimated but also compared the ICCs obtained for different rating scales. To do this, we estimated all ICCs for each relevant effect of the 3×2 factorial design imposed by the manipulation of the rating scales. More specifically, we estimated the ICCs for the pooled rating scales, for each scale format separately, for discrete/continuous scales separately, and for each rating scale separately.

Once these ICCs were estimated, we investigated main and interaction effects of the factorial design by pairwise comparison. We did so by subtracting the posterior distributions of the ICCs of two different conditions and computing the 95% credible interval (CrI) for the resulting distribution. We then assessed whether 0 was a part of this interval and inferred that if this was not the case, the ICCs were different in value. This analysis allowed us to assess the main and interaction effects of format and continuity on the consistency of the rating scale as a whole.

In keeping with what was explained above, we compared the ICCs in two ways. First, we compared the ICCs associated with the valence ratings, allowing us to compare all rating scale formats to each other. Second, we compared the ICCs associated with the PA and NA ratings, which limited us to a comparison of the ESG and the 2D VAS only.

Questionnaire

In the second part of our analyses, we delved deeper into the questionnaire responses. Like the consistency analyses, these analyses consisted of two parts: The estimation of scores for each assessed variable and the comparison of these variables across conditions.

First, we performed a confirmatory principal component analysis (PCA) on the questionnaire data using the *psych* package in R (Revelle, 2022). This analysis allowed us to group all items of the questionnaire into different components. We specified four components to be found with an oblique rotation, allowing correlations between the components. After running the PCA, we computed component scores using the same package. Given that we allowed for correlations between the components, this approach to computing component scores is warranted.

Then, we compared the component scores in a similar way as the ICCs. More specifically, we used a nonparametric bootstrap to generate 10,000 new samples by randomly drawing participants from the observed sample. Within each of the generated samples, we preserved the observed unequal sample sizes across conditions. More concretely, we sampled the same number of participants from each condition to match the observed number of individuals for that specific condition. In other words, all drawn samples had a sample size per condition equal to those shown in Table 1.

After resampling, we computed the mean of the component scores for each drawn sample according to the 3 × 2 factorial design. Just as for the ICCs, we computed the means for the pooled scales, the different formats, the discrete/continuous scales, and all scales separately. Finally, we did pairwise comparisons between the conditions by subtracting two sampled distributions at a time and by computing the 95% CI of these distributions. We thus use the same procedure to establish the presence of the main and/or interaction effects for the questionnaire data as we did for the ICCs.

Deviation from preregistration

Here, we detail how we deviated from the preregistration, and more importantly, why.

First, we changed some aspects of the estimation code for the estimation of the ICCs. More specifically, we removed the bounds on the main and interaction effect parameters in Eqs. 1 and 5. Originally, these parameters were bound between zero and half of the range of affect, following other researchers (ten Hove et al., 2022a). However, we found that imposing these bounds led to convergence issues and estimates that did not match frequentist estimates of the ICCs. After the change, the chains converged and the Bayesian and frequentist estimates were similar in value.

Another change to the estimation of the ICCs concerns the priors. Originally, we planned on using a very broad uniform distribution as a prior on the variance parameters based on the results of a simulation study. However, after imposing the change in bounds, we switched to a more appropriate prior, namely the lognormal distribution.

A third change occurred in the questionnaire analysis. In the code, we originally computed the component scores by multiplying the questionnaire responses with the component loadings of each item. However, we later opted for the use of a function provided by the *psych* package to do this computation for us.

Finally, there were several nonsignificant changes to the code that can be found on GitLab. Most of these changes were fixes of small mistakes in the preprocessing code (e.g., the first author forgot to include *format* and *continuity* as variables in the dataset after preprocessing the data), fixes

Table 3 Summary statistics for the raw dependent variables of interest in this study. These statistics are shown for each of the scales separately. Additionally, the average within-person correlations between the affective variables and the monetary outcomes of the experiment are provided (r_{outcome})

Format	Continuity	Variable	*M*	*SD*	Min	Max	Skew	Kurtosis	$r_{outcome}$
1D VAS	Discrete	Valence	0.48	0.24	0.11	0.87	−0.01	−0.76	.74 (*SD* =.25)
	Continuous	Valence	0.50	0.23	0.11	0.90	−0.05	−0.55	.71 (*SD* =.26)
ESG	Discrete	Valence	0.48	0.20	0.16	0.83	0.04	−0.58	.68 (*SD* =.26)
		PA	0.43	0.23	0.13	0.83	0.29	0.63	.64 (*SD* =.26)
		NA	0.48	0.23	0.14	0.84	0.11	−0.30	−.57 (*SD* =.29)
	Continuous	Valence	0.49	0.23	0.10	0.89	0.00	−0.93	.75 (*SD* =.21)
		PA	0.46	0.26	0.06	0.91	0.16	−0.86	.70 (*SD* =.23)
		NA	0.47	0.24	0.07	0.90	0.11	−0.67	−.62 (*SD* =.28)
2D VAS	Discrete	Valence	0.49	0.20	0.16	0.83	−0.01	−0.03	.69 (*SD* =.26)
		PA	0.40	0.22	0.13	0.81	0.54	1.16	.66 (*SD* =.26)
		NA	0.43	0.22	0.13	0.83	0.38	0.88	−.58 (*SD* =.27)
	Continuous	Valence	0.48	0.21	0.11	0.89	0.07	−0.46	.68 (*SD* =.27)
		PA	0.42	0.24	0.07	0.89	0.40	0.09	.64 (*SD* =.26)
		NA	0.45	0.23	0.06	0.90	0.22	−0.27	−.56 (*SD* =.28)

of technical issues (e.g., imposed a check on whether we had all data for a given participant), and additions of data and results to the repository.

Results

Description of affective data

Table 3 describes the characteristics of the affective measurements for each of the scales separately. The average within-person correlation between affect and the experimental stimuli is furthermore included (r_{outcome}). Turning first to the general descriptives, one may notice that there are no big differences between the scales with regard to mean, standard deviation, and range. With regard to skewness and kurtosis, the measurements do show some differences, with the most obvious one being a more positive kurtosis for PA and NA as measured by the 2D VAS compared to PA and NA as measured by the ESG. This suggests that the 2D VAS might elicit more evenly distributed (i.e., less peaked) responses than the other formats.

Turning to the correlations, one can see that measured affect was strongly related to the experimental stimuli. This suggests that participants' affective states were influenced by the wins and losses that they encountered in the experiment.

Description of parameters

Table 4 shows the mean and standard deviation for the estimated values of the parameters and the proportion of variance explained R^2 of the moderated autoregressive model defined in Eq. 4. We highlight two observations in this table. First, parameter estimates are not too different across the different affective variables. This suggests that in this experiment our conclusions would be very similar across the different variables, such that both the monetary outcomes and the previous affective state matter for capturing the affective fluctuations observed in this experiment. Second, the R^2 is relatively high, showing that these two predictors indeed capture much of the observed variance in the affective variables.

Consistency and agreement

Descriptive results

Table 5 summarizes the values of the ICCs at the group level, providing the reader with a general overview of the consistencies found in our experiment. As can be seen in this table, the ICC_{person} was fairly low, suggesting that the individuals' mean affective states could not be adequately distinguished in the affective data. Although slightly higher, the $ICC_{outcome}$ was also not very high, ranging from poor to moderate consistency. These results indicate that neither the individual differences nor the systematic effect of the monetary outcomes could be adequately picked up in our measurements.

One reason for these results might be the fact that only mean tendencies were accounted for in the computation of these ICCs. More specifically, the ICC_{person} examines how much of the observed variance could be attributed to individual differences in mean affective states. However, while some individual differences might be expected in the participants' trial-by-trial use of the scale, these differences were averaged

Table 4 Mean and standard deviations for the estimated values for the parameters of the moderated autoregressive model for each of the conditions. Additionally, the mean and standard deviation of the R^2 for the model is also shown. Values for the parameters are very similar across the different affective variables and the R^2 indicates that a good portion of the variance is explained by the predictor variables, namely the monetary outcomes and the previous indicated affective state

	Intercept (β_0)		Outcome (β_1)		Autoregression (β_2)		R^2	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Valence	0.35	0.16	0.56	0.31	0.27	0.28	0.74	0.20
Positive affect	0.31	0.15	0.58	0.32	0.24	0.26	0.67	0.21
Negative affect	0.33	0.16	−0.49	0.31	0.29	0.28	0.60	0.22

Table 5 Summary of the posterior distribution for the ICCs across all participants and rating scales. ICCs ranged from moderate to good, with the lowest consistency being that of general individual differences in affect (ICC_{person}) and the highest being the individual differences in the relation between affect and the stimuli of the experiment ($ICC_{1,\,person}$)

	ICC_{person}		$ICC_{outcome}$		$ICC_{0,\,person}$		$ICC_{1,\,person}$		$ICC_{2,\,person}$	
	M	95% *CrI*	*M*	95% *CrI*	*M*	95% *CrI*	*M*	95% *CrI*	*M*	95% *CrI*
Valence	0.34	[0.32,0.36]	0.58	[0.42,0.75]	0.70	[0.40,0.77]	0.84	[0.72,0.88]	0.62	[0.36,0.68]
PA	0.39	[0.36,0.41]	0.55	[0.39,0.73]	0.72	[0.48,0.77]	0.84	[0.70,0.87]	0.57	[0.36,0.63]
NA	0.37	[0.35,0.40]	0.46	[0.30,0.65]	0.53	[0.23,0.63]	0.81	[0.70,0.84]	0.52	[0.36,0.57]

Note: For the parameter-based ICCs the *j* refers to the intercept when $j = 0$, to the slope for the monetary outcomes when $j = 1$, or to the autoregressive effect when $j = 2$

out by our analysis. This made it unlikely that individual differences would be picked up by this ICC. A similar case can be made for the $ICC_{outcome}$, which examines how consistently specific sizes of the monetary outcomes go together with a given affective state. This explanation is corroborated by a quick examination of the $\sigma^2_{person\times outcome}$, which quantifies the differences in how people reacted to the experimental stimuli (see Eqs. 1-2). Following suggestions from the literature (e.g., Liljequist et al., 2019; ten Hove et al., 2022b), we treated these individual differences as error in our estimation routine. However, this variance takes a big chunk of the complete variance cake (posterior $M_{\sigma^2} \in [\,0.11, 0.12]$, or about half of the total variance; see Table 3), evidently leading to a significant reduction in both the ICC_{person} and $ICC_{outcome}$.

Turning to the parameter-based ICCs, one may notice that all were found to be within the range of moderate to good consistency. Interestingly, the agreement of the slope of the monetary outcomes on affect was especially high (values between 0.81 and 0.84), indicating that individual differences in affective reactivity were consistently picked up by the scales. Agreement was worst for the autoregressive effect, for which the ICCs fell in the "moderate" area (values between 0.52 and 0.62). This result indicates that the extent to which current affective states depended on previous affective states was not adequately picked up. In other words, while participants remained consistent in their response to the experimental stimuli ($ICC_{1,\,person}$), they did not consistently take their previous answers into account.

In sum, we found poor to good consistency of the repeated affective measurements. Importantly, the size of the ICC depended on whether we decomposed the raw affective data (ICC_{person} and $ICC_{outcome}$) or specific parameters of interest from a statistical model ($ICC_{j,\,person}$), with the former ICCs being poor and the latter moderate to good.

Comparison

Tables detailing the results of the pairwise comparisons can be found in the *Supplementary Materials*. Summarizing the results, we found only very few differences between the rating scales: Only for the ICC_{person} did we find a main effect of format when comparing the ESG to the 2D VAS and an interaction effect when comparing the 1D VAS to the 2D VAS. With regard to the former effect, we found that the 2D VAS had a consistently higher ICC_{person} than the ESG across all affective variables ($M_{difference} \in [0.11,0.14]$). Regarding the latter effect, we found that the 2D VAS almost always had a higher ICC_{person} than the 1D VAS ($M_{difference} \in [0.09,0.13]$) with exception of the comparison between the discrete 2D VAS and the continuous 1D VAS, which was almost marked as different ($M_{difference} = 0.06$, $95\%_{difference} = [-0.01,0.12]$). None of the other ICCs differed across rating scales.

Summarizing these results, we found little evidence for the notion that the scales under investigation differ with respect to the consistency of their measurements.

Questionnaires

Principal component analysis

Detailed descriptive statistics for the questionnaire responses and the result of the PCA can be found in the *Supplementary Materials*. The component structure that we found in the PCA did not align perfectly with the presumed structure in Table 2. Similar to the presumed structure, we found a component measuring the ease with which participants used the rating scales (Component 2, hereafter *Ease*) and the accuracy with which they described their feelings using the scales (Component 3, hereafter *Accuracy*). These two components showed overlap in one of the items, namely item 14: "I was able to accurately describe my feelings with the emotion measure."

Contrary to our expectations, we found two new components that rearranged the items of *Manipulation check* and *Motivation*. The first component combined all items of *Motivation* with the first two questions of *Manipulation check* ("The experiment elicited positive/negative feelings"). We will therefore refer to this component as the *Experience* component, detailing how participants experienced the experiment. The second component combined two negative feelings items ("The experiment elicited negative feelings" and "The experiment was frustrating") with the two items that measured indifference to the experiment ("I felt indifferent about the experiment" and "Winning or losing money did not have any effect on me"). Because both clusters had opposite loadings, and because the indifference items loaded positively on this component, we will refer to this component as *Indifference*.

Comparison

In the *Supplementary Materials*, the interested reader can find detailed tables of the pairwise comparisons of the component scores across scales. We found a significant main effect of *format* for both *Ease* and *Accuracy*. For *Ease*, the 1D VAS was rated as easier to use than the 2D VAS, which in turn was easier to use than the ESG ($M_{1D\ VAS} = 0.44$, $M_{2D\ VAS} = 0.03$, $M_{ESG} = -0.46$). For *Accuracy*, we found that the 1D VAS was rated as more accurate than the 2D VAS ($M_{1D\ VAS} = 0.16$, $M_{2D\ VAS} = -0.09$, $M_{difference} = 0.25$).

We furthermore found a significant interaction effect of *format* and *continuity* for *Accuracy*. This interaction effect entailed a significant difference between the 1D VAS and the ESG ($M_{difference} \in [0.18, 0.28]$), with exception of the comparison between the discrete 1D VAS and the discrete ESG ($M_{difference} = 0.16$, $95\%_{difference} = [-0.02, 0.34]$). This result suggests that the 1D VAS was perceived as more accurate than the ESG, except when participants were restricted to discrete responding. No other significant differences were found.

Summarizing these results, we found that the 1D VAS was perceived as being the most user-friendly and most accurate of the scale formats. Thus, while consistency did not differ across scales, participants' perception of the scales differed in favor of the simpler, one-dimensional scale.

Discussion

In this study, we investigated the consistency of several ratings scales that are regularly used in experimental settings. We found poor to good consistency of the measurements obtained with the rating scales, depending on the way in which the ICCs were computed. It is important to note that the two kinds of ICCs we computed in this study seemingly contradicted each other. More specifically, we found poor to moderate consistency of the participant and monetary outcome effects (ICC_{person} and $ICC_{outcome}$) and moderate to good agreement for the estimated values of the parameters of a linear model ($ICC_{j,\ person}$). On the one hand, the ICCs showed that the scales did not pick up individual differences in mean affective responding (ICC_{person} was low), but on the other hand, we did find that a participant's estimated parameters of a linear model remained roughly the same across the experiment ($ICC_{j,\ person}$ were high). This contradiction can be explained by the nature of our paradigm, given that the crux of this task is the variability in one's responses in relation to the monetary outcomes, which is explicitly modeled for the parameter-based ICCs, but is averaged out in the affect-based ones.

Our results show that the consistency of the measure depends on the research question that is asked. Within this experimental setting, for example, one might opt to use the raw affective measurements to study individual differences in mean levels of valence, PA, or NA. However, the researcher who does so might be disappointed with the result, as individual differences in mean responses could not be adequately distinguished in these data. If, on the other hand, one would study these same individual differences with respect to one's reactivity to the stimuli of the experiment, one might get a more robust result. Our consistency study thus revealed that the probabilistic reward paradigm can be used safely to investigate individual differences in model-based affect dynamics, given that these individual differences can be sufficiently well distinguished. However, using the raw affect ratings to differentiate between people (or stimuli for that matter) based on their average level is not a good idea because of poor consistency.

Aggregating across all ICCs, our results suggest that at least 16% of the variation in affective measurements is due to measurement error. These results are in line with the findings of daily-life studies (see, e.g., Dejonckheere et al., 2022; Eisele et al., 2021; Wilhelm & Schoebi, 2007). Given that the

consistencies found in this study also showed a large range of values depending on the exact variable of interest, we conclude that researchers should mind their measurements, even within a controlled experimental environment. For this, researchers can follow our example by conducting one or more pilot studies in which they assess the consistency of their measurements. If the consistency is deemed unsatisfactory, researchers can then increase the strength of the manipulation to improve consistency, though this may introduce other potential issues (e.g., the manipulation may not hold up to ethical standards, Lord et al., 1968). Alternatively, researchers can use analysis techniques like structural equation models and use the estimated error variance from the pilot study to account for errors in their subsequent study (Anderson & Gerbing, 1988).

When comparing the ICCs across the rating scales, we found little to no relationship between the ICCs and the scale's format or continuity. This suggests that the type of scale did not influence consistency, meaning that participants used each kind of scale in an equally consistent way. This null result is nonetheless interesting for two reasons. First, the different scale formats varied in complexity. This was corroborated by the finding that participants experienced the two-dimensional rating scales as the most difficult to use and the least accurate for capturing their feelings. Even so, this added complexity did not lead to a decreased consistency of their ratings. This might suggest that even despite some initial confusion about the meaning of a scale (see the *Supplementary Materials*), participants eventually learned to use the scales consistently.

A second reason why our null results are interesting lies in the fact that it goes against the commonly held assumption that a decreased number of response alternatives should lead to increased consistency in responses (see, e.g., Krosnick & Fabrigar, 1997). Some studies challenged this hypothesis, suggesting that the choice between these two types of scales is not important with regard to consistency (for a short overview, see Aguinis et al., 2009). Our findings are in line with these latter studies. However, our study suffers from the fact that we used a rather large number of discrete responses of our scales (9) compared to the typically used number of response options (e.g., 5), which might have unintentionally occluded a difference in consistency between discrete and continuous responses. Future research should therefore aim to more closely examine the effect of scale coarseness on the consistency of the scale.

With regard to the questionnaire, we found that participants subjectively preferred the 1D VAS with regard to the ease and accuracy with which they used this scale. Given that consistency was not influenced by the format or the continuity of the scales, our results might serve as a motivation for researchers to use the 1D VAS in their own research.

As with any study, there are some limitations to consider. First, the results of this study are inherently linked to the probabilistic reward task that we used. The consistency that we found in this simple task—in which participants were repeatedly exposed to monetary outcomes—might not necessarily generalize to other types of experimental designs such as mood-induction paradigms. Future research might therefore benefit from investigating the consistency of affective ratings scales in different types of designs, giving us an overview of the consistency within each type of design.

Second, we focused on the comparison of several affective rating scales while keeping other variables constant. It is, however, conceivable that other experimental manipulations might influence the consistency of these measures. For example, the number of trials a participant went through was kept constant in this study, but it is possible that a higher number of trials leads to greater frustration at the end of the experiment, and thus to decreased consistency in a participants' affective responses. While we did not observe such a frustration effect in our study (see the *Supplementary Materials*), future research should shed light on whether these kinds of experimental manipulations do indeed impact the consistency of one's measurements.

Third, participants were deceived in our study, as they were presented with the same sequence of monetary outcomes repeatedly during the experiment. In our analyses, we have assumed that participants did not notice this manipulation—an assumption that was based on a pilot study in which we found that participants were not aware of the repeated sequences of outcomes. However, it is possible that some participants noticed the fixed nature of the monetary outcomes, which might have significantly influenced our results. We are unable to rule out this possibility because we did not incorporate a question about participants' suspicions regarding this issue. This limitation should be remediated in future studies.

Finally, we investigated the consistency of our scales by using a test–retest experimental design. However, this type of analysis assumes that participants' affective lives do not change in the meantime, that is, that participants affectively respond in the same way throughout the experiment. This type of assumption might not necessarily hold, even in a strictly controlled experiment like this one. More concretely, one could make the argument that participants' affective states are an accumulation of all previously encountered stimuli in the experiment, meaning that no one sequence of monetary outcomes in our experiment is ever the same as another. If this is indeed the case, then some of the systematic affective fluctuations still seeped into our estimate of the measurement error. Future research might therefore benefit from trying to assess consistency in diverse ways within the same experimental design, for example through the additional application of model-based approaches (Schuurman et al., 2015).

Acknowledgements We would like to thank Peter Kuppens for his help at the early stages of this study. We would furthermore like to thank Nena Lathouwers, who helped us with executing a pilot study and analyzed the data that came out of it. We would also like to thank Kenny Yu for giving his opinion on earlier drafts of this paper. The analyses performed in this work were performed using resources and services of the VSC (Flemish Supercomputer Center), funded by the Research Foundation – Flanders (FWO) and the Flemish Government.

Open practice statement In line with suggestions of the Open Science movement, we preregistered this study. This preregistration can be found on the Open Science Framework: https://osf.io/sytrn. We furthermore preregistered code, which can be found under the *preregistered* tag on the Gitlab page of this study. As stated earlier, data and materials can also be found on this same page.

Code availability Participants can find the code for the analyses on the same GitLab page, repeated here: https://gitlab.kuleuven.be/ppw-okpiv/researchers/u0123135/affective-consistency.

Author contribution NV and FT conceptualized the study together. NV conceptualized and performed the analyses with valuable help from SV, AM, WV, and FT. NV, SV, AM, WV, and FT all wrote and reviewed the article.

Data availability The interested reader can find the data and experimental code on the Gitlab repository of this study: https://gitlab.kuleuven.be/ppw-okpiv/researchers/u0123135/affective-consistency.

Declarations

Conflict of interest The authors declare no conflict of interest.

Funding This work was supported by the Research Fund of the KU Leuven (Grant C14/19/054) and by the Fonds Wetenschappelijk Onderzoek (FWO; Grant G074219N). The funders had no role in study design, data collection, analyses, decision to publish, or preparation of the manuscript.

Ethical approval As stated in the article, this study was approved by the local ethics committee at the Psychological department of the KU Leuven (the Social and Societal Ethics Committee) under case number G-2021-3228. The study was performed in accordance with the ethical standards as laid out in the 1964 Declaration of Helsinki.

Consent to participate As stated in the article, participants signed an informed consent before participating in our study.

Consent to publish Within the informed consent, participants were informed on our intention to publish the results of the study. Participants consented to the submission of this study's results for publication.

References

Adolf, J. K., Voelkle, M. C., Brose, A., & Schmiedek, F. (2017). Capturing context-related change in emotional dynamics via fixed moderated time series analysis. *Multivariate Behavioral Research, 52*, 499–531. https://doi.org/10.1080/00273171.2017.1321978

Aguinis, H., Pierce, C. A., & Culpepper, S. A. (2009). Scale coarseness as a methodological artifact: Correcting correlation coefficients attenuated from using coarse scales. *Organizational Research Methods, 12*, 623–652. https://doi.org/10.1177/1094428108318065

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*, 411–423.

Asutay, E., Genevsky, A., Feldman-Barrett, L., Hamilton, J. P., Slovic, P., & Västfjäll, D. (2021). Affective calculus: The construction of affect through information integration over time. *Emotion, 21*, 159–174. https://doi.org/10.1037/emo0000681

Bonsall, M. B., Wallace-Hadrill, S. M. A., Geddes, J. R., Goodwin, G. M., & Holmes, E. A. (2012). Nonlinear time-series approaches in characterizing mood stability and mood instability in bipolar disorder. *Proceedings of the Royal Society B, 279*, 916–924. https://doi.org/10.1098/rspb.2011.1246

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag.

Bulteel, K., Mesdagh, M., Tuerlinckx, F., & Ceulemans, E. (2018). VAR(1) based models do not always outpredict AR(1) models in typical psychological applications. *Psychological Methods, 23*, 740–756. https://doi.org/10.1037/met0000178

Burns, R. A., & Ma, J. (2015). Examining the association between psychological wellbeing with daily and intra-individual variation in subjective wellbeing. *Personality and Individual Differences, 82*, 34–39. https://doi.org/10.1016/j.paid.2015.02.023

Cunningham, W. A., Dunfield, K. A., & Stillman, P. E. (2013). Emotional states from affective dynamics. *Emotion Review, 5*, 344–355. https://doi.org/10.1177/1754073913489749

Dejonckheere, E., Demeyer, F., Guesens, B., Piot, M., Tuerlinckx, F., Verdonck, S., & Mestdagh, M. (2022). Assessing the reliability of single-item momentary affective measurements in experience sampling. *Psychological Assessment, 34*, 1138–1154. https://doi.org/10.1037/pas0001178

Dejonckheere, E., Houben, M., Schat, E., Ceulemans, E., & Kuppens, P. (2021). The short-term psychological impact of the COVID-19 pandemic in psychiatric patients: Evidence for differential emotion and symptom trajectories in Belgium. *Psychologica Belgica, 61*, 163–172. https://doi.org/10.5334/pb.1028

Dejonckheere, E., & Mestdagh, M. (2021). On the signal-to-noise ratio in real-life emotional time series. In C. E. Waugh & P. Kuppens (Eds.), *Affect dynamics*. Springer. https://doi.org/10.1007/978-3-030-82965-0_7

Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature: Human. Behaviour, 3*, 478–491. https://doi.org/10.1038/s41562-019-0555-0

Driver, C. C., & Voelkle, M. C. (2018). Understanding the time course of interventions with continuous time dynamic models. In K. van Montfort, J. H. L. Oud, & M. C. Voelkle (Eds.), *Continuous time modeling in the behavioral and related sciences*. Springer.

Eisele, G., Lafit, G., Vachon, H., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2021). Affective structure, measurement invariance, and reliability across different experience sampling protocols. *Journal of Research in Personality, 92*, 104094. https://doi.org/10.1016/j.jrp.2021.104094

Eldar, E., & Niv, Y. (2015). Interaction between emotional state and learning underlies mood instability. *Nature Communications, 6*, 6149. https://doi.org/10.1038/ncomms7149

Frijda, N. H. (2007). *The laws of emotion*. Routledge.

Haney, A. M., Fleming, M. N., Wycoff, A. M., Griffin, S. A., & Trull, T. (2023). Measuring affect in daily life: A multilevel psychometric evaluation of the PANAS-X across four ecological momentary assessment samples. *Psychological Assessment*. https://doi.org/10.1037/pas0001231

Huys, Q. J. M., Pizzagalli, D. A., Bogdan, R., & Dayan, P. (2013). Mapping anhedonia onto reinforcement learning: A behavioural meta-analysis. *Biology of Mood & Anxiety Disorders, 3*. https://doi.org/10.1186/2045-5380-3-12

Kalokerinos, E. K., Murphy, S. C., Koval, P., Bailen, N. H., Crombez, G., Hollenstein, T., Gleeson, J., Thompson, R. J., Van Ryckeghem, D. M. L., Kuppens, P., & Bastian, B. (2020). Neuroticism may not reflect emotional variability. *Proceedings of the National Academy of Science, 117*, 9270–9276. https://doi.org/10.1073/pnas.1919934117

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality*. John Wiley & Sons, Inc.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* ((2nd ed.). ed.). Academic Press.

Kuppens, P., & Verduyn, P. (2017). *Emotion dynamics. Current Opinion in Psychology, 17*, 22–26. https://doi.org/10.1016/j.copsyc.2017.06.004

Larsen, J. T., Norris, C. J., McGraw, A. P., Hawkley, L. C., & Cacioppo, J. T. (2008). The evaluative space grid: A single-item measure of positivity and negativity. *Cognition & Emotion, 23*(3), 453–480. https://doi.org/10.1080/02699930801994054

Liljequist, D., Elfving, B., & Roaldsen, K. S. (2019). Intraclass correlation – A discussion and demonstration of basic features. *PLoS ONE, 14*, e0219854. https://doi.org/10.1371/journal.pone.0219854

Loossens, T., Mestdagh, M., Dejonckheere, E., Kuppens, P., Tuerlinckx, F., & Verdonck, S. (2020). The Affective Ising Model: A computational account of human affect dynamics. *PLoS Computational Biology, 16*, e1007860. https://doi.org/10.1371/journal.pcbi.1007860

Loossens, T., Tuerlinckx, F., & Verdonck, S. (2021). A comparison of continuous and discrete time modeling of affective processes in terms of predictive accuracy. *Scientific Reports, 11*, 6218. https://doi.org/10.1038/s41598-021-85320-4

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lucas, R. E., & Donnellan, M. B. (2012). Estimating the reliability of single-item life satisfaction measures: Results from four national panel studies. *Social Indicators Research, 105*, 323–331. https://doi.org/10.1007/s11205-011-9783-z

Matheson, G. J. (2019). We need to talk about reliability: Making better use of test-retest studies for study design and interpretation. *PeerJ, e6918*. https://doi.org/10.7717/peerj.6918

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46.

Moors, A., Van de Cruys, S., & Pourtois, G. (2021). Comparison of the determinants for positive and negative affect proposed by appraisal theories, goal-directed theories, and predictive processing theories. *Current Opinion In Behavioral Sciences, 39*, 147–152. https://doi.org/10.1016/j.cobeha.2021.03.015

Polit, D. F. (2014). Getting serious about test–retest reliability: A critique of retest research and some recommendations. *Quality of Life Research, 23*, 1713–1720. https://doi.org/10.1007/s11136-014-0632-9

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385–401. https://doi.org/10.1177/014662167700100306

Revelle, W. (2022). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.R-project.org/package=psych

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences, 111*(33), 12252–12257. https://doi.org/10.1073/pnas.1407535111

Schuurman, N. K., Houtveen, J. H., & Hamaker, E. L. (2015). Incorporating measurement error in n = 1 psychological autoregressive models. *Frontiers in Psychology, 6*, 1038. https://doi.org/10.3389/fpsyg.2015.01038

Scott, S. B., Sliwinski, M. J., Zawadzki, M., Stawski, R. S., Kim, J., Marcusson-Clavertz, D., Lanza, S. T., Conroy, D. E., Buxton, O., Almeida, D. M., & Smyth, J. M. (2020). A coordinated analysis of variance in affect in daily life. *Assessment, 27*, 1683–1698. https://doi.org/10.1177/1073191118799460

Smillie, L. D., Geaney, J. T., Wilt, J., Cooper, A. J., & Revelle, W. (2013). Aspects of extraversion are unrelated to pleasant affective reactivity: Further examination of the affective-reactivity hypothesis. *Journal of Research in Personality, 47*, 580–587. https://doi.org/10.1016/j.jmp.2013.04.008

Stan Development Team. (2022). *RStan: The R interface to Stan*. https://mc-stan.org/

ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2022a). Interrater reliability for multilevel data: A generalizability theory approach. *Psychological Methods, 27*, 650–666. https://doi.org/10.1037/met0000391

ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2022b). Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. *Psychological Methods.* https://doi.org/10.1037/met0000516

Trull, T. J., Lane, S. P., Koval, P., & Ebner-Priemer, U. W. (2015). Affective dynamics in psychopathology. *Emotion Review, 7*, 355–361. https://doi.org/10.1177/1754073915590617

Vanhasbroeck, N., Ariens, S., Tuerlinckx, F., & Loossens, T. (2021). Computational models for affect dynamics. In C. E. Waugh & P. Kuppens (Eds.), *Affect dynamics*. Springer.

Vanhasbroeck, N., Loossens, T., Anarat, N., Ariens, S., Vanpaemel, W., Moors, A., & Tuerlinckx, F. (2022). Stimulus-driven affective change: Evaluating computational models of affect dynamics in conjunction with input. *Affective Science, 3*, 559–576. https://doi.org/10.1007/s42761-022-00118-5

Villano, W. J., Otto, A. R., Ezie, C. E. C., Gillis, R., & Heller, A. S. (2020). Temporal dynamics of real-world emotions are more strongly linked to prediction error than outcome. *Journal of Experimental Psychology: General, 149*, 1755–1766. https://doi.org/10.1037/xge0000740

Wendt, L. P., Wright, A. G. C., Pilkonis, P. A., Woods, W. C., Denissen, J. J. A., Kühnel, A., & Zimmermann, J. (2020). Indicators of affect dynamics: Structure, reliability, and personality correlates. *European Journal of Personality, 34*, 1060–1072. https://doi.org/10.1002/per.2277

Wilhelm, P., & Schoebi, D. (2007). Assessing mood in daily life: Structural validity, sensitivity to change, and reliability of a short-scale to measure three basic dimensions of mood. *European Journal of Psychological Assessment, 23*, 258–267. https://doi.org/10.1027/1015-5759.23.4.258